PA153 Počítačové zpracování přirozeného jazyka

06 – Korpusy a korpusové nástroje, značkování

Karel Pala, Vít Suchomel

Centrum ZPJ, FI MU, Brno

21. října 2013

Definice

Korpus je soubor dat (textů) v přirozeném jazyce.

Použití

- obecně: data ke studiu přirozeného jazyka
- lexikografové: slovníky
- lingvisté: jazykové analýzy, změny jazyka
- sociologové: jak a o čem píšeme, která témata jsou aktuální
- marketingoví experti: hodnocení značek a výrobků v textech
- statistické nástroje ZPJ: jazykové modely pro značkovače, analyzátory, překladové systémy, prediktivní psaní,. . .

Příklady zdrojů dat

- tištěná média: knihy, časopisy, noviny, básně
- internet: články, prezentace, blogy, diskuze, tweety
- řeč: přepis záznamů řeči, filmové titulky
- ostatní: osobní korespondence, školní eseje

Zvláštní vlastnosti korpusů

- podle data vzniku obsahu: synchronní x diachronní
- jednojazyčné x vícejazyčné
- srovnatelné x paralelní
- podle zkrácení dokumentů: plné texty x zkrácené vzorky
- média: audio (záznam dialogu), video (záznam emocí)

Tradiční textové korpusy

Vznik

- obvykle na objednávku vládní instituce, univerzity nebo nakladatelství
- zdroje: obvykle z tištěných médií – nakladatelství, skenování knih, přepisy rozhovorů

Výhody tradičních korpusů

- kontrolovaný obsah (vyvážená reprezentace žánrů a stylů)
- kvalitní a bohaté informace o datech (autor, název, rok vydání, žánr, styl, oblast)
- možnost opravy chyb

Nevýhody tradičních korpusů

- nedostatenčná velikost pro některá použití
- obtížné získávání dat, vysoké náklady
- problémy s autorskými právy

Standard Corpus of Present-Day American English (Brown corpus)

- Brown University (Henry Kucera, W. Nelson Francis)
- 1964 (1971, 1979)
- 500 vzorků textu délky 2000 slov každý = 1 mil. slov
- `http://khnt.aksis.uib.no/icame/manuals/brown/`

British National Corpus (BNC)

- Oxford University, Longman
- 1991–1994 (2001, 2007)
- vzorky textu délky 100 mil. slov dohromady
- 90 % psaná řeč, 10 % mluvená řeč
- `http://www.natcorp.ox.ac.uk/`

Corpus of Contemporary American English (COCA)

- Brigham Young University (Mark Davies)
- od 1990, každý rok přidáno 20 mil. slov
- 450 mil. slov (2013)
- `http://corpus.byu.edu/coca/`

Český národní korpus SYN

- Ústav ČNK na FF UK v Praze
- texty od 1990 vydání SYN2000, SYN2005, SYN2010
- 1,3 mld. slov (2010)
- `http://korpus.cz/`

Korpus DESAM

- CZPJ FI MU
- morfologicky označkovaný korpus českých textů
- desambiguované (jednoznačné) značkování
- 1 mil. slov

Web je největší korpus

Myšlenka a iniciativa „Web as Corpus“ (`http://sigwac.org.uk/`)
Výhody internetových korpusů

- obrovské množství dat
- dokumenty různých druhů
- aktuální podoba psané formy jazyka
- snadná dostupnost, nízké náklady

Nevýhody internetových korpusů

- neuspořádanost
- nežádoucí obsah
- duplicity
- chyby
- víme, co stahujeme?

Proč potřebujeme velké korpusy?

Přínosy velkých korpusů

- větší slovník (více různých slov)
- více/lepší příklady použití slov ve větách
- lepší pokrytí řídkých jazykových jevů
- více dat pro přesnější jazykové modely

Velké textové korpusy získané z internetu v CZPJ

jazyk	velikost korpusu [GB]	velikost korpusu [10^9 tokenů]	doba stahování [dny]
enTenTen12	108	17.8	17
esAmTenTen11	44	8.7	14
arTenTen12	58	6.6	28
czTenTen11		5.8	40
frTenTen12	72	12.4	15
jpTenTen11	61	11.1	28
ruTenTen12	198	20.2	14
turecké texty	26	4.1	14

V NLPC máme k dispozici také kolekci dat ClueWeb ‘09 — vyčištěná anglická část obsahuje zhruba 70 miliard tokenů.

Paralelní korpus InterCorp

- Ústav ČNK na FF UK v Praze
- jazykové páry (vždy s češtinou) zarovnané na větách
- 10–30 mil. slov každý pár
- `http://korpus.cz/intercorp/`

Další paralelní korpusy

- OPUS – dosptupná paralelní data (`http://opus.lingfil.uu.se/`)
- Europarl – jednání EP (`http://www.statmt.org/europarl/`)
- 1984 – Orwellův román (`http://nl.ijs.si/ME/Vault/CD/docs/1984.html`)

Google Books Ngrams

- Vyhledávání ve skenovaných knihách
- Pouze ntice slov ($n \in \{1..5\}$)
- `https://books.google.com/ngrams`

Postup získávání webových korpusů v CZPJ

- příprava jayzkově závislých modelů používaných v dalších krocích — učení na dokumentech z Wikipedie
- spuštění crawleru (`SpiderLing`)
- zpracování a vyhodnocování během běhu crawleru
  - detekce znakové sady dokumentu (`Chared`)
  - filtrování jazyka (vektor trigramů znaků)
  - odstraňování nežádoucího obsahu (`Justext`)
  - kontrola duplicitních dokumentů
  - vyhodnocování průběžné výtěžnosti webových domén
- zpracování získaných dat
  - odstranění podobných odstavců (`Onion`)
  - tokenizace (`Unitok` nebo jiný nástroj)
  - značkování morfologické a syntaktické — externími nástroji, jsou-li dostupné
  - zakódování a nahrání do korpusového manažeru (`Manatee`/`Bonito`)

Více v předmětu PA154 nástroje pro korpusy

Web crawler

Web crawler je druh počítačového programu

- prochází internet (stránky propojené odkazy)
- stahuje dokumenty (metainformace, obsah)
- ukládá části dokumentů v různých formátech k dalšímu použití

Crawlery

- k získávání obsahu dokumentů – GoogleBot (navíc k indexování), Heritrix a mnoho dalších
- ke sbírání odkazů
- k získávání textových dokumentů pro ZPJ – SpiderLing

Ukázka dat v korpusu – XML vertikální formát

```
<dokument zanr="blog"
  nazev="Dovolená v Paříži" datum="2011-10-28"
  url="http://karel.bloguje.cz/dovolena-v-parizi">
<odstavec nadpis="1">
<veta>
Po
sedmi
letech
v
kouzelné
Paříži
!
</veta>
</odstavec>
...
</dokument>
```

Obecný korpusový manažer

- příprava textu – převod z různých formátů
- zahrnutí metadat (informací o datech – zdroj, autor, téma, žánr, datum)
- tokenizace (rozdělení na slova, interpunkce, znaky)
- anotace (značkování)
- efektivní uchování korpusu – datové struktury umožňující rychlé získání uložených dat
- konkordance – získání úseků textů odpovídajících uživatelským dotazům
- výpočet statistik – vyhledání typických vzorů v datech, frekvenční distribuce, souvýskyty

Word Sketch Engine

- korpusový manažer (a více)
- vyvíjený od roku 2000 v CZPJ FI MU (dizertační práce Pavla Rychlého)
- od 2003 spolupráce s průmyslovým partnerem Lexical Computing
- hlavní komponenty
  - Manatee – korpusový manažer
  - Bonito – uživatelské rozhraní a API
  - Corpus Architect – vytváření uživatelských korpusů a jejich nahrávání do Manatee
- pro zaměstnance a studenty MU zdarma na `https://ske.fi.muni.cz`

Manatee – korpusový manažer

- akceptuje XML vertikální formát dat
- podporuje metadata a anotace, jsou-li správně předzpracovány
- korpusy uchovává efektivně
- konkordance – získání úseků textů odpovídajících uživatelským dotazům
- Word Sketch = slovní profil – stručný přehled kolokačního a gramatického chování slova
- výpočet statistik – vyhledání typických vzorů v datech, frekvenční distribuce, souvýskyty
- *více v předmětu PA154 Statistické nástroje pro korpusy (jaro 2014)*

Corpus Query Language (CQL)

- dotazovací jazyk podporovaný Manatee
- slouží k vyhledání tokenů v korpuse
- využívá regulárních výrazů
- příklad: `[lemma="červený"|lemma="černý"] [tag="k1.*nP.*"]`
  dvě bezprostředně následující slova, první má základní tvar ,,červený“ nebo ,,černý“, druhé je podstatné jméno v množném čísle, například ,,červenými domky“ je platný odpovídající výraz

Bonito – uživatelské rozhraní a API

- převádí uživatelské dotazy do CQL
- volá funkce Manatee
- výsledek zobrazuje uživateli nebo ve formátu JSON pro API
- ukázka: `https://ske.fi.muni.cz`

Corpus Architect – uživatelské korpusy

- zajišťuje autentizaci a přístup uživatelů k jejich korpusům
- ukládá a zpracovává uživatelská data
- zpracovaná data nahrává do Manatee
- obsahuje univerzální tokenizaci
- pracuje s morfologickými analyzátory pro více než 10 jazyků
- zahrnuje nástroj WebBootCaT k získávání korpusů z internetu

Alternativy k některým funkcím Sketch Engine

- samostatné vyhledávací nástroje pro daný korpus (např. BNC)
- WordSmith (Mike Scott, `http://www.lexically.net/wordsmith`)
- AntConc (Laurence Anthony, `http://www.antlab.sci.waseda.ac.jp/antconc_index.html`)

Anotace

Anotace je přidávání lingvistických informací do korpusu.

- informace o zpracování dat (např. rozdělení na tokeny)
- metadata textů (zdroj, autor, téma, žánr, datum)
- struktury (dokument, odstavec, věta, zarovnání, mluvčí)
- značkování – přiřazení značky (např. slovního druhu) k tokenu

Druhy anotace

- morfologická (slovní druh a jiné gramatické kategorie)
  - u nás (čeština): morfologický analyzátor Majka
  - jiné: TreeTagger (enTenTen12), CLAWS (BNC, COCA), FreeLing (esTenTen11)
- syntaktická (parsing – závislostní nebo složkové stromy, chunking – rozdělení na fráze jmennou /NP/, slovesnou /VP/, předložkovou /PP/)
  - u nás (čeština): Synt, SET, DIS/VADIS, IOBBER (polština)
  - jiné: MST Parser, MaltParser
- sémantická (word sense tagging/desambiguation /WSD/ – rozlišení významu slova, named entity recognition – rozpoznání jmenných entit /NER/)
  - u nás (čeština): DESAMB – desambiguace morfologických značek
  - jiné: WordNet, SuperSenseTagger – WSD, NER
- koreference (určení anafory)
  - u nás (angličtina): SARA
- pragmatická (označení mluvčího, komunikační situace)

Ukázka anotací v korpusu – XML vertikální formát

```
<dokument zanr="blog" nazev="Dovolená v Paříži">
<veta nadpis="1">
Po        po        k7c6           0  8
sedmi     sedm      k4c6           1  7
letech    léto      k1gNnPc6       2  7
v         v         k7c6           3 10
kouzelné  kouzelný  k2eAgFnSc6d1   4  9
<entita druh="město">
Paříži    Paříž     k1gFnSc6       5  9
</entita>
!         !         kx             6 11
<NP>                               7  8
<PP>                               8 11
<NP>                               9 10
<PP>                              10 11
<S>                               11  -
</veta>
```

Editory anotací

- výstup vždy v XML
- GATE `http://gate.ac.uk/`
- Brat `http://brat.nlplab.org/`
- WordSmith `http://www.lexically.net/wordsmith`
- u nás: Phrase Annotator (shallow parsing: fráze, závislosti), Sysel (sémantické kategorie)

Problémy s anotacemi

Manuální x automatická

- Ruční anotace je zdlouhavá a nákladná. Přesto nemusí být dokonalá.
- Nedokonalá automatická anotace (naučená na ručně anotovaných datech) je pro velká data nevyhnutelná.

Cyklické anotace (podle lekce Corpus Mark-up)

- Data v korpusu pozorujeme skrz anotace. Byly-li kategorie anotací zvoleny a anotace provedena ještě před průzkumem korpusu, došlo k omezení předem, na jaké otázky se můžeme při pozorování korpusu ptát.
- Řešením je cyklicky
  - analyzovat korpus
  - na základě toho volit parametry anotací
  - anotace provádět

Literatura

- Kilgarriff, Adam, Gregory Grefenstette. Introduction to the special issue on the web as corpus. In Computational linguistics 29.3 (2003): s. 333-347.
- RYCHLÝ, Pavel a Pavel SMRŽ. Manatee, Bonito and Word Sketches for Czech. In Proceedings of the Second International Conference on Corpus Linguisitcs. Saint-Petersburg: Saint-Petersburg State University Press, 2004. s. 124-132, 9 s.
- KILGARRIFF, Adam, Pavel RYCHLÝ, Pavel SMRŽ a David TUGWELL. The Sketch Engine. In Proceedings of the Eleventh EURALEX International Congress. Lorient, France: Universite de Bretagne-Sud, 2004. s. 105-116, 12 s.
- Corpus Query Language ve Sketch Engine: `http://trac.sketchengine.co.uk/wiki/SkE/CorpusQuerying`
- Lekce Corpus Mark-up od Matthew Brook O'Donnela z UoL Summer Institute in Corpus Linguistics: `www.lexically.net/courses/sessions/markup/Corpus%20Mark-up.ppt`